A Guide to Using NCCS JAGUAR System

OAK RIDGE LEADERSHIP COMPUTING FACILITY

U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Outline

- Jaguar System Overview
- Logging into Jaguar
- Login Nodes vs. Compute Nodes
- File Systems
- Software Environment
- Compiling
- Running Jobs
- Third-Party Software
- Resources for Users

Outline: Jaguar System Overview

– Summary of Resources

– System Software

– System Hardware

Jaguar System Overview: Summary of Resources

Jaguar is a Cray XT system consisting of XT4 and XT5 partitions!

Jaguar	XT4	XT5
CPU Type	2.1 GHz Quad-core AMD Opteron (Budapest)	2.6 GHz Hex-core AMD Opteron (Istanbul)
Interconnect	Cray SeaStar2 Router	Cray SeaStar2+ Router
Switching Capacity (Router's Peak Bandwidth)	45.6GB/s 6 switch ports per Cray SeaStar chip, 7.6 GB/s each	57.6GB/s 6 switch ports per Cray SeaStar2+ chip, 9.6 GB/s each
Memory type	DDR2-800 (some nodes use DDR2-667 memory)	DDR2-800
Memory Bandwidth	10.6 to 12.8 GB/sec per AMD Opteron	21.2 GB/sec to 25.6 GB/sec per compute node
Floor Space	1400 feet^2	4400 feet^2
Cooling Technology	Air	Liquid

Jaguar System Overview: Summary of Resources

Jaguar is a Cray XT system consisting of XT4 and XT5 partitions!

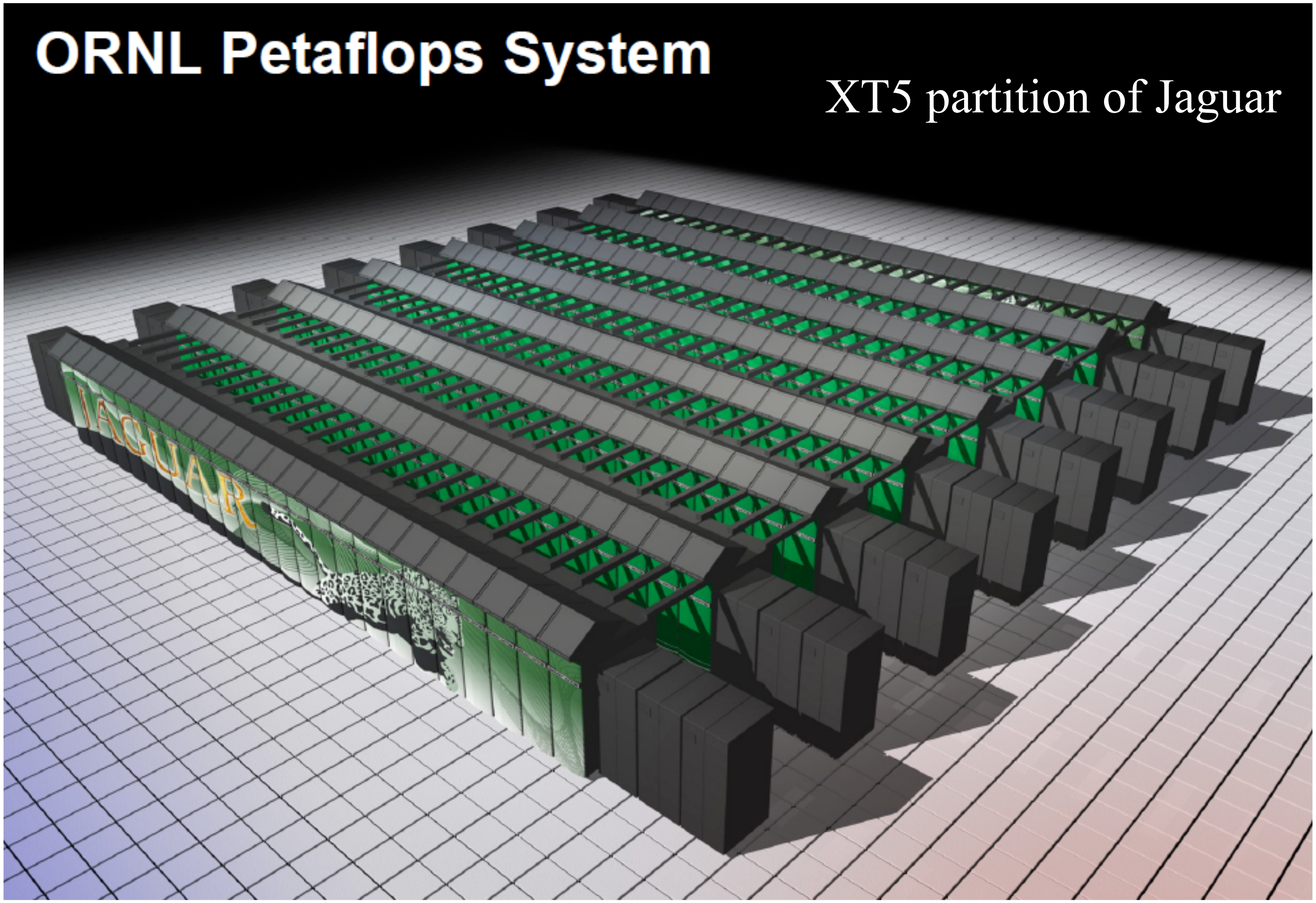
Jaguar	XT4	XT5	Total
Nodes per blade	4		
CPUs per node[1]	1	2	
Cores per node	4	12	
Compute nodes[2]	7,832	18,688	
AMD Opteron cores	31,328	224,256	255,584
Memory per CPU	8 GB/CPU		
System Memory	~61.2 TB	~292 TB	~353.2 TB
Disk Bandwidth	~44 GB/s	~240 GB/s	~284 GB/s
Disk Space	~750 TB	~10,000 TB	~10,750 TB
Interconnect Bandwidth	~157 TB/s	~374 TB/s	~532 TB/s
Floor Space	1400 $feet^2$	4400 $feet^2$	5800 $feet^2$
Ideal Performance per core[3] (4 FLOPs/cycle times $2.1*10^9$ cycles/sec)	8.4 GFLOPS	10.4 GFLOPS	
Overall Ideal Performance	~263.16 TFLOPS	~2.33 PFLOPS	~2.60 PFLOPS

[1] In the context of Jaguar CPU is also called a socket.
[2] Note that in addition to compute nodes Jaguar also has input/output (I/O) and login service nodes.
[3] FLOPs = **FL**oating point **OP**eration**s**; FLOPS = **FL**oating point **O**perations **P**er **S**econd

Jaguar System Overview: System Software

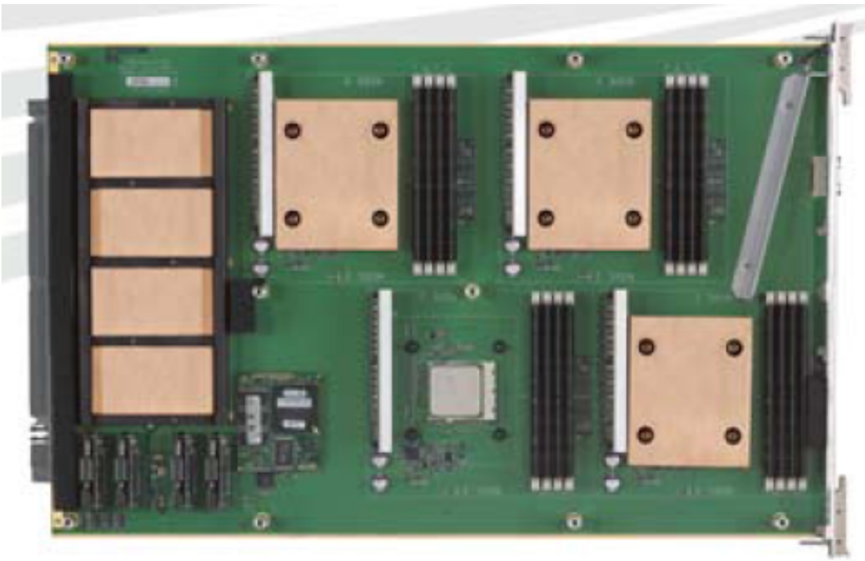
- Operating system is Cray Linux Environment (CLE) 2.1:
  - Compute Nodes – Compute Node Linux (CNL)
  - Login/Service nodes – SUSE Linux
- Compilers
  - C/C++, Fortran
- MPI implementation
  - Cray MPI based on MPICH
- High Performance Storage System (HPSS) software

Jaguar System Overview: System Hardware



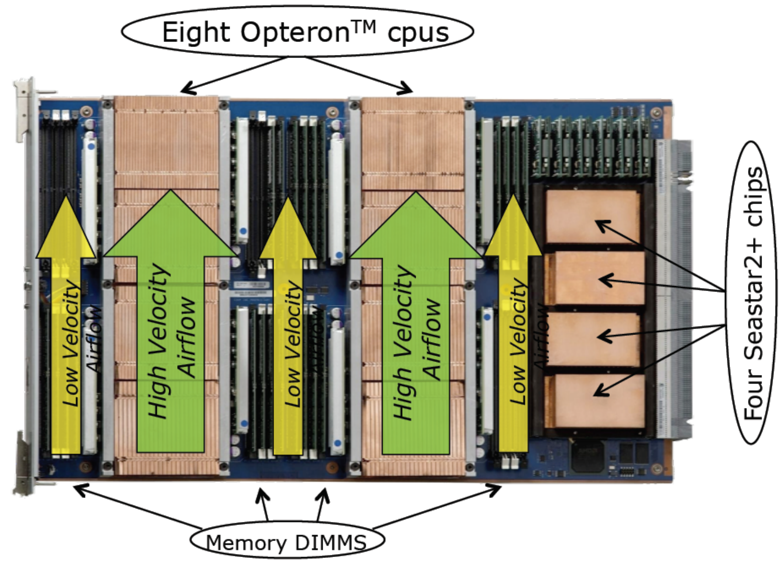
Jaguar System Overview: System Hardware

Cray XT4 Blade



Four Compute Nodes per Blade
ONE AMD quad-core CPU per node

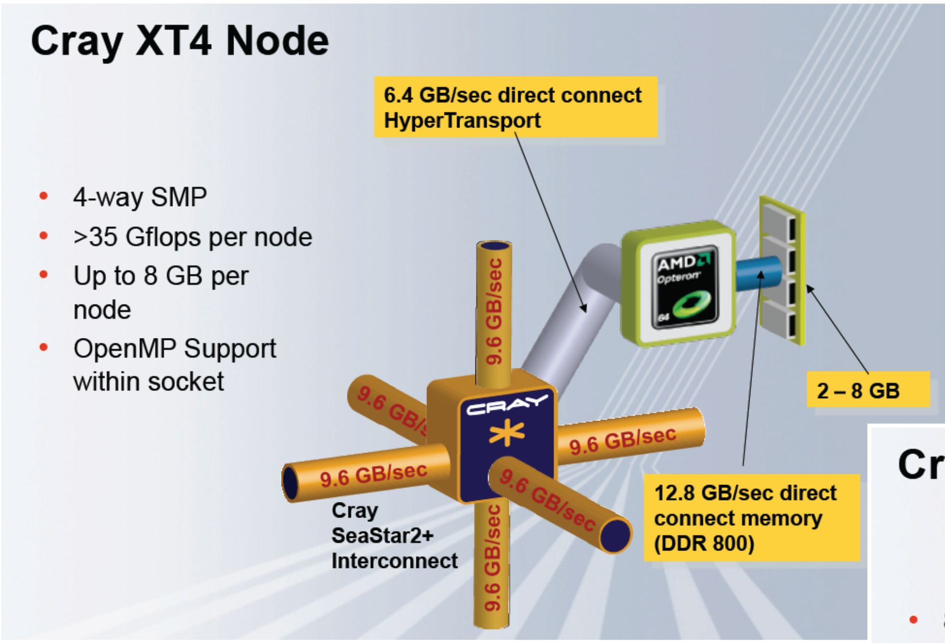
Cray XT5 Blade



Four Compute Nodes per Blade
TWO AMD hex-core CPUs per node

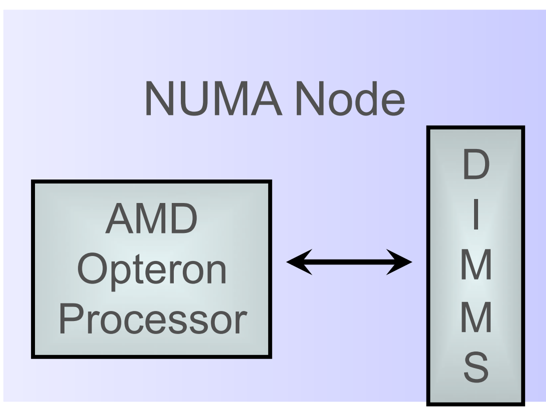
Go to Menu

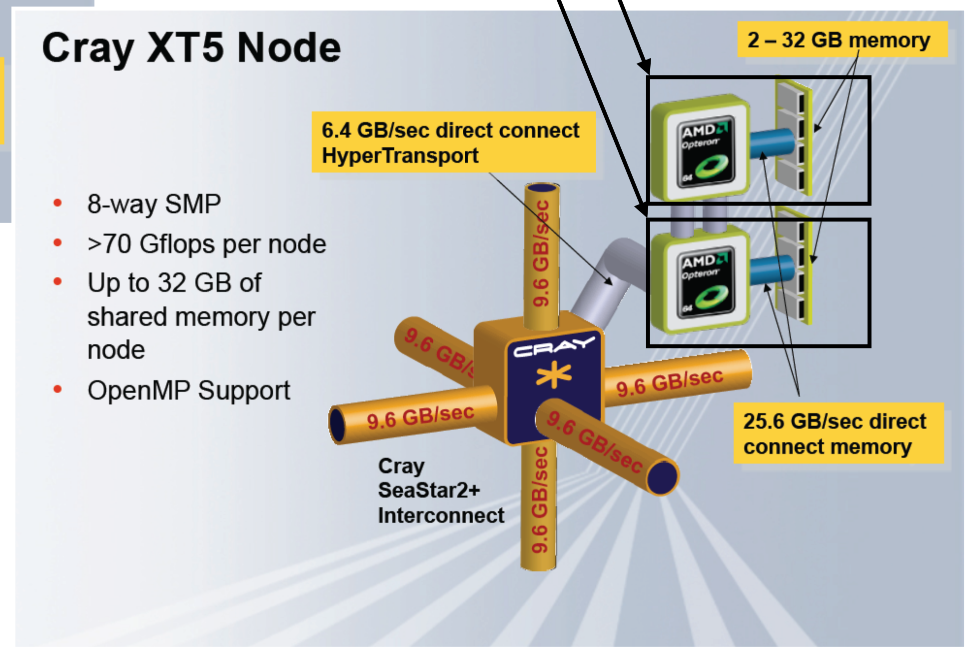
Jaguar System Overview: System Hardware



NUMA Node 0

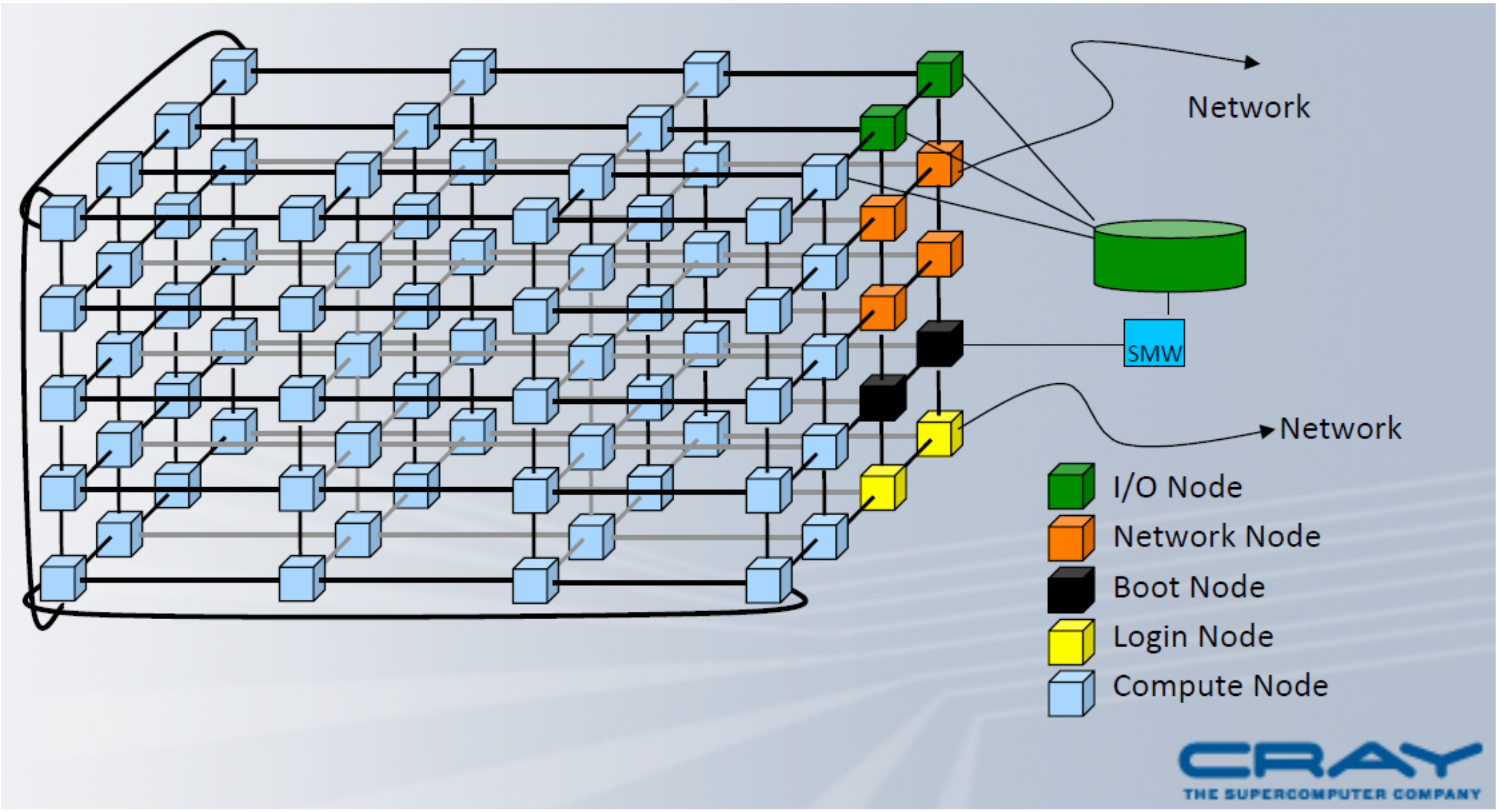
NUMA Node 1





Jaguar System Overview: System Hardware

- XT System Torus Architecture



Go to Menu

Jaguar System Overview: System Hardware

Cache Hierarchy
Quad-core AMD Opteron CPU

Dedicated L1 cache

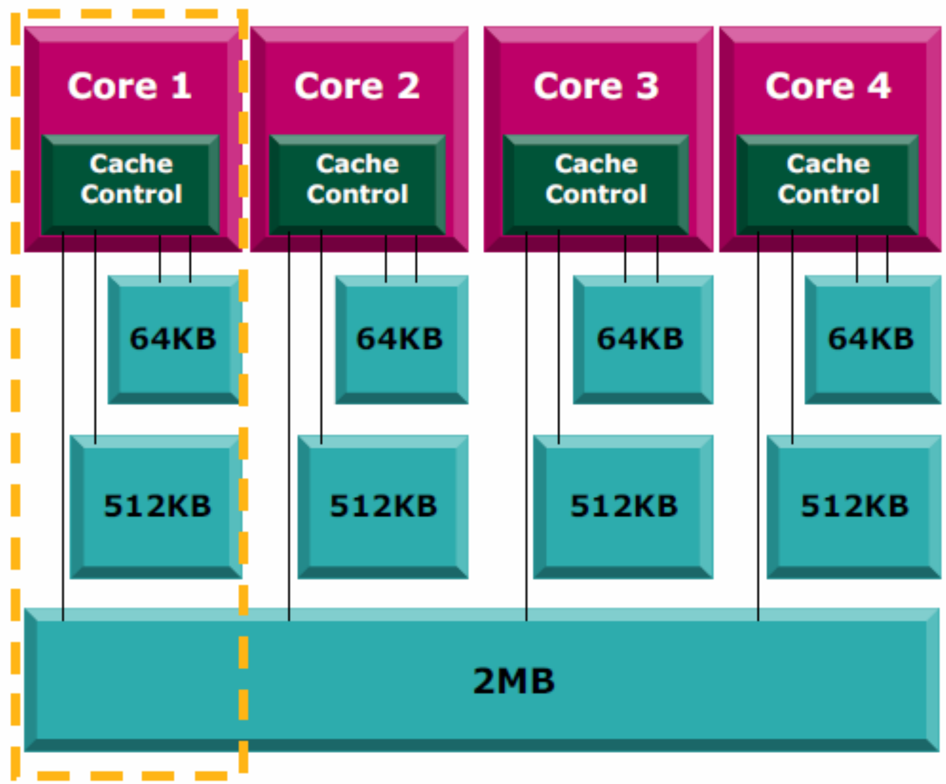
- 2 way associativity.
- 8 banks.
- 2 128bit loads per cycle.

Dedicated L2 cache

- 16 way associativity.

Shared L3 cache

- fills from L3 leave likely shared lines in L3.
- sharing aware replacement policy.



Outline: Logging into Jaguar

- Connection Requirements
- Connection Procedures
- One-Time Password (OTP) Authentication
- Connection Options

Logging into Jaguar: Connection Requirements

- The only supported remote client on NCCS systems is a secure shell (SSH) client.
- The only supported authentication method is one-time passwords (OTP).

- UNIX-based operating systems generally have an SSH client built in.
- Windows users may obtain free clients online, such as PuTTY.

Any SSH client:

- must support the SSH-2 protocol (supported by all modern SSH clients).
- must allow keyboard-interactive authentication to access NCCS systems. For UNIX-based SSH clients, the following line should be in either the default ssh_config file or your $HOME/.ssh/config file:

```
    PreferredAuthentications keyboard-
interactive,password
```

The line may also contain other authentication methods, so long as keyboard-interactive is included.

Logging into Jaguar: Connection Procedures

One more time: Jaguar is a combined system of Cray XT4 and XT5 systems!

To connect to Jaguar from a UNIX-based system type the following in your terminal:

```
ssh userid@jaguar.ccs.ornl.gov     - Cray XT4
ssh userid@jaguarpf.ccs.ornl.gov   - Cray XT5
```

Enter PASSCODE: PIN + 6 digits from RSA® SecurID

NCCS RSA Key Fingerprints:

```
jaguar       0d:c9:db:37:55:da:41:26:55:4a:80:bb:71:55:dd:01
jaguarpf     80:58:21:03:96:47:1a:15:2c:25:d3:ca:e6:04:e8:a7
```

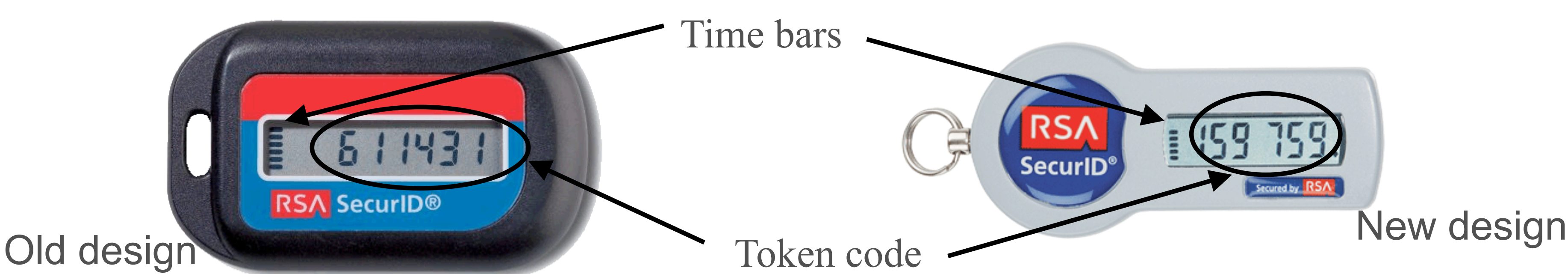
Logging into Jaguar: One-Time Password (OTP) Authentication

RSA® SecurID - Quick Start Guide

All NCCS systems currently use OTPs as their authentication method. To login to NCCS systems, an RSA SecurID key fob is required!

Activating your SecurID key fob:

1. Return the completed NCCS token activation form to the address provided on the form. Once the form is received by NCCS, the RSA OTP token will be enabled, and you will be notified by email.
2. Initiate an SSH connection to home.ccs.ornl.gov
3. When prompted for a PASSCODE, enter the token code shown on the fob. You will be asked if you are ready to set your PIN. Answer with "Y."
4. You will then be prompted to enter a PIN. Enter a 4- to 6-digit number you can remember. Reenter your PIN when prompted.
5. You will be prompted to enter your full PASSCODE. To do so, wait until the next token code appears on your fob and enter your PASSCODE, which is now your PIN + the 6-digit token code displayed on your fob. For example, if the pin was 1111 and the token code was 223344, then the full PASSCODE would be 1111223344.
6. Your PIN is now set, and your fob is activated and ready for use. Log on using the procedure outlined in



Logging into Jaguar: Connection Options

X11 Tunneling

Automatic forwarding of the X11 display to a remote computer is highly recommended with the use of SSH and a local X server. To set up an automatic X11 tunneling with SSH, do one of the following:

1. Command line: Invoke ssh with the -X option:

   ssh -X userid@jaguarpf.ccs.ornl.gov

   Note 1: use of the -x (lowercase x) option will disable X11 forwarding
   Note 2: use of the -Y option (instead of -X) is necessary on some systems to enable "trusted" X11 forwarding

2. Configuration file: Edit (or create) the .ssh/config file to have the following line in it:

   ForwardX11 yes

3. Graphical Menu: Many SSH clients have a menu to change the configuration settings.
   PuTTY: check the box next to Connection --> SSH --> X11 --> Enable X11 Forwarding
   Note 1: Unix-like systems, with the exception of Mac OS-X, offer native X11 support. Apple does provide an implementation for OS-X, available from the Apple website.
   Note 2: For Windows systems you can also use free Xming software.
   Note 3: PuTTY stores configuration settings for each server separately. If, for example, you enable X11 Forwarding for `jaguar.ccs.ornl.gov`, it will not change the settings for `jaguarpf.ccs.ornl.gov`

Outline: Login Nodes vs. Compute Nodes

– Login Nodes

– Compute (Batch) Nodes

Login Nodes

- When you login to Jaguar, you will be placed on a "login node"
- Login nodes are used for basic tasks such as file editing, code compilation, data backup, and job submission
- These nodes provide a full SUSE Linux environment, complete with compilers, tools, and libraries
- The login nodes should not be used to run production jobs. Production work should be performed on the systems compute resources.
- Serial jobs (post-processing, *etc*) may be run on the compute nodes as long as they are statically linked (will be discussed later)

Compute (Batch) Nodes

- All MPI/OpenMP user applications execute on batch or compute nodes

- Batch nodes provide **limited** Linux environment – Compute Node Linux (CNL)

- Compute nodes can see only the Lustre scratch directories

- Access to compute resources is managed by the PBS/TORQUE – batch system manager

- Job scheduling is handled by Moab, which interacts with PBS/TORQUE and the XT system software.

Outline: File Systems

- Basics
- User’s Directories
- High Performance Storage System (HPSS)
- Lustre Filesystem

File Systems: Basics

- The Network File Service (NFS) server contains user's home directories, project directories, and software directories .
- Compute nodes can only see the Lustre work space
  - The NFS-mounted home, project, and software directories are not accessible to the compute nodes.
- Shared Lustre area (SPIDER) is now available on compute nodes and is the only scratch area for the XT5.
- Executables must be executed from within the Lustre work space:
  - `/tmp/work/$USER` (XT4 and XT5)
  - `/lustre/scr144/$USER` (XT4 only)
- Batch jobs can be submitted from the home or work space. If submitted from a user’s home area, a batch script should cd into the Lustre work space directory (`cd $PBS_O_WORKDIR`) prior to running the executable through aprun.
- All input must reside in the Lustre work space
- All output must also be sent to the Lustre work space

File Systems: User's Directories

Each user is provided the following space resources:

- Home directory - NFS Filesystem
  `/ccs/home/$USER`

- Work directory/Scratch space - Lustre Filesystem
  `/tmp/work/$USER`

- Project directory - NFS Filesystem
  `/ccs/proj/projectid`

- HPSS storage

File Systems: Home Directory

- Each user is provided a home directory to store frequently used items such as source code, binaries, and scripts. Home directories are located in a Network File Service (NFS) that is accessible from all NCCS resources.

- Home directory - NFS Filesystem

  Location: `/ccs/home/$USER`

- Accessible from all NCCS systems
- NFS does not provide the highest performance
- Default storage limit of 2 GB
- To find your quota and usage in NFS, use the `quota` command
- Regularly backed up

File Systems: Work Directory

- Work space is available on each NCCS high-performance computing (HPC) system for temporary files and for staging large files from and to the High Performance Storage System (HPSS). To ensure adequate work space is available for user's jobs, a script that finds and deletes old files runs on the system nightly. Thus, it is critical to archive files from the scratch area as soon as possible.
- Work directory/Scratch space - Lustre Filesystem
  - `/tmp/work/$USER` (both XT4 and XT5)
  - `/lustre/scr144/$USER` (local, XT4 only)
- The path `/tmp/work/$USER` is available on all NCCS HPC systems.
- On the NCCS Cray XT4 (jaguar), Cray XT5 (jaguarpf), Data Analysis Cluster (lens), and Development Cluster (smoky), the /tmp/work/$USER path points to the same shared lustre area. On the IBM Blue Gene/P (eugene) the link points to a local area within the system's General Parallel File System (GPFS).
- **Not backed up!**

File Systems: Project Directory

- Each project is provided a directory shared by the project to store data such as source code, binaries, and scripts. Project directories are located in a Network File Service (NFS) that is accessible from all NCCS resources.

- Project directory - NFS Filesystem

  Location: `/ccs/proj/projectid`

- Accessible from all NCCS systems
- Default storage limit of 5 GB
- By default, project directories are created with 770 permissions and the project ID group as the group owner.

File Systems: Node-local System

Node-local Filesystem

- Location: `/tmp`

- The path `/tmp/work/$USER` is available on all NCCS HPC systems, but it points to a different file system.

- Do not create files directly in the `/tmp` directory!

- The `/tmp` file system itself is quite small, and when `/tmp` fills up, the system problems result.

File Systems: High Performance Storage System (HPSS)

- HPSS is an archival Back-up system which consists of
  - two types of storage technology:
    - disk – "on-line" for frequently/recently accessed files
    - tape – "off-line" for very large or infrequently accessed files
  - Linux servers
  - High Performance Storage System software
- Tape storage is provided by robotic tape libraries.

- HPSS has three SL8500 tape libraries. Each can hold up to 10,000 cartridges.
- The StorageTek SL8500 libraries house a total of
  - twenty-four T10000A tape drives (500 gigabyte cartridges, uncompressed)
  - thirty-six T10000B tape drives (1 terabyte cartridges, uncompressed).
- Each drive has a bandwidth of 120 MB/s
- As of October, 2009, HPSS has 7.2 PB stored in over 16.1 million files.

File Systems: Using `hsi` and `htar` on HPSS

- Each user of an NCCS system is provided an account on the HPSS. The user's login name for HPSS is the same as for all other NCCS systems. Authorization to HPSS is by means of the user's SecurID token.
- **Users are encouraged to use `hsi` when dealing with a small number of files, and `htar` for large numbers of files.**
- The `hsi` utility provides the ability to access and transfer data to and from the NCCS HPSS for both disk and tape file systems. Issuing the command `hsi` will start HSI in interactive mode.
- Information on HSI may be found from the NCCS systems through the command
  - `hsi help`
- The `htar` command – works like Unix "tar"
- Below is an example of storing and getting a bunch of files in a directory using tar and HSI. HSI can read from standard input and write to standard output:
  - `tar cvf - . | hsi put - : <filename.tar>`
  - `hsi get - : <filename.tar> | tar xvf -`

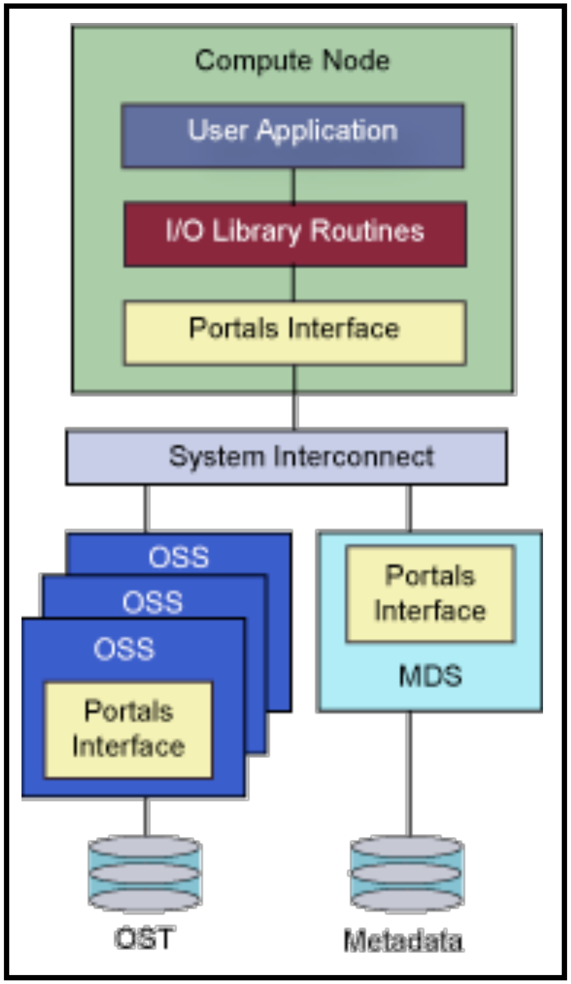
File Systems: Center-wide File System (SPIDER)

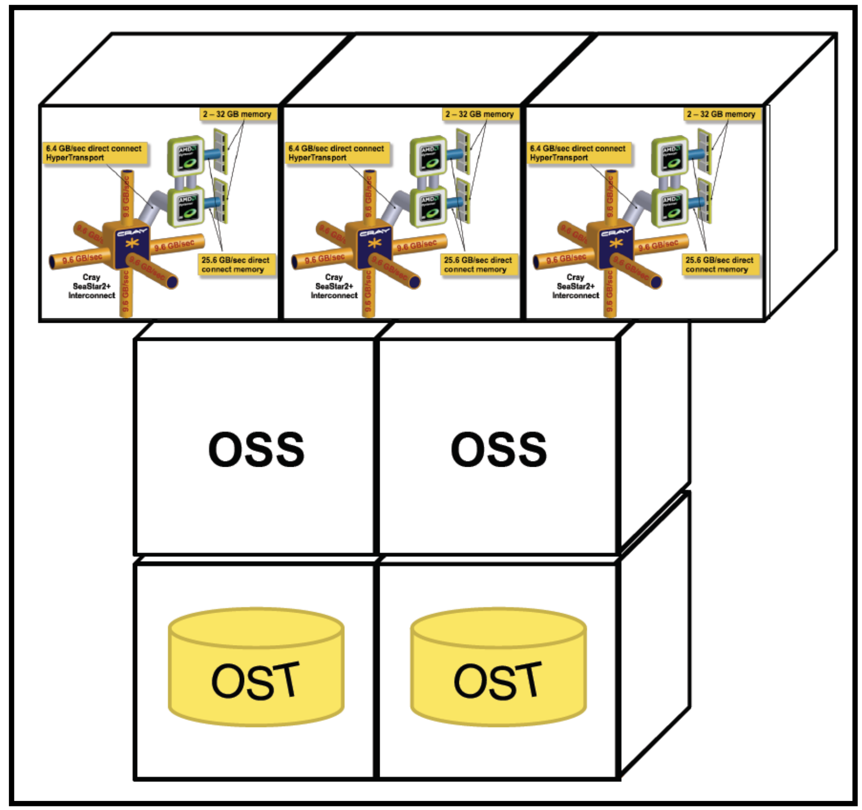
- "Spider" provides a shared, parallel file system for all LCF systems
  - Based on Lustre file system
- Over 10 PB of RAID-6 Capacity
  - 13,440 1Gb SATA Drives (10,652 for production and 2,688 for parity)
    - 48 caplets * 28 tiers/caplet * 10 discs/tier = 13,440 SATA discs
    - 33 tons of discs:-)
  - 192 Open Storage Servers (OSS) and 1344 Object Storage Targets (OST)
    - 192 OSSs * 7 OSTs/OSS = 1344 OSTs
  - 3 Terabytes of memory
- Demonstrated bandwidth of over 200 GB/s (244 Gb/s aggregate R/W)
  - 30,000 files created per second
- Demonstrated stability on a number of LCF Systems
  - Over 26,000 lustre clients at NCCS mounting the file system and performing I/O
- Available from all systems via our high performance scalable I/O network
  - 4 InfiniBand core switches
  - Over 3,000 InfiniBand ports
  - Over 3 miles of cables

File Systems: Lustre Filesystem and `liblustre`

- Parallel, object-based filesystem that aggregates a number of storage servers together to form a single coherent file system that can be accessed by a client system. The Lustre file system is made up of an underlying set of file systems called Object Storage Targets (OST's), which are essentially a set of parallel IO servers.
- When running on compute node, only Lustre filesystem is accesible
  - Exception: stdin, stdout and stderr are mapped through aprun
- Best way to do I/O on compute nodes *without going back out through aprun* (thereby throttling I/O badly): using `liblustre`
- Lustre module is currently loaded by default
  - Linked in when you build executable
- Do not unload Lustre module.
- Use the ftn, cc, and CC wrappers to compile

File Systems: Lustre Filesystem - a Bigger Picture





File Systems: Striping on Lustre Filesystem

- A file is said to be striped when read and write operations access multiple OST's concurrently. File striping is a way to increase IO performance since writing or reading from multiple OST's simultaneously increases the available I/O bandwidth
- Striping will likely have little impact for the following IO patterns:
  – Serial IO where a single processor or node performs all of the IO for an application.
  – Multiple nodes perform IO, access files at different times.
  – Multiple nodes perform IO simultaneously to different files that are small (each < 100 MB).
- You can change the striping pattern across the OSTs on a *per directory* basis yourself
  – Default stripe width is 4
- You should have a good understanding of how and how much your application outputs before you attempt this!

Lustre Filesystem: Striping on Lustre Filesystem (cont…)

- You should think of striping as "preparing the ground for I/O."
  - The striping occurs the next time you write to the directory/file
  - If you change the settings for an existing directory, you will need to copy the files elsewhere and then copy them back to inherit the new settings
- `lfs getstripe` *`filename`* will tell you the striping information for a file
  - `lfs find -v <dir/file>` is equivalent
- `lfs setstripe <dir>` *`size start number`*
  - Defaults: `-s 1M -c 4 -i -1`
  - `lfs setstripe <dir>` *`0 -1 1`* means no striping
  - Caution: *You can fill up individual OSTs!*

• `lfs setstripe`	`lfs setstripe`
– Stripe size	– s (default:1M, k, M, G)
– Stripe count	– c 5 (default 4, -1 All)
– Stripe index	– i 0 (default: -1 round robin)
	`<file	directory>`

Outline: Software Environment

– Defaults

– Modules

– module command

Software Environment: Defaults

- Default software environment automatically loaded when user logs in (Nov. 2009):
  - PGI 9.0.4
  - Libsci 10.4.0
  - MPT 3.5.0
  - Etc.

- What if…
  - Not all software necessary for your work is automatically loaded
  - You need another version of loaded software
  - Default software incompatible with your work?

Software Environment: Modules

- Software is loaded, unloaded or swapped using modules.
- Use of modules allows software, libraries, paths, etc. to be cleanly entered into and removed from your programming environment.
- Conflicts are detected and module loads that would cause conflicts are not allowed.

Software Environment: `module` command

Loading Commands

- **`module [load||unload] my_module`**
  - Loads/Unloads module *`my_module`*
  - e.g., `module load subversion`
- **`module swap module#1 module#2`**
  - Replaces *`module#1`* with *`module#2`*
  - e.g., `module swap PrgEnv-pgi PrgEnv-gnu`

Informational Commands

- **`module help my_module`**
  - Lists available commands and usage
- **`module show my_module`**
  - Displays the actions upon loading `my_module`
- **`module list`**
  - Lists all loaded modules
- **`module avail [name]`**
  - Lists all modules [beginning with *`name`*]
  - e.g., `module avail gcc`

Software Environment: `module list`

```
username@jaguarpf-login1:/> module list
Currently Loaded Modulefiles:
  1) modules/3.1.6
  2) DefApps
  3) torque/2.4.1b1-snap.200905191614
  4) moab/5.3.6
  5) /opt/cray/xt-asyncpe/default/modulefiles/xtpe-istanbul
  6) cray/MySQL/5.0.64-1.0000.2342.16.1
  7) xtpe-target-cnl
  8) xt-service/2.2.41A
  9) xt-os/2.2.41A
 10) xt-boot/2.2.41A
 11) xt-lustre-ss/2.2.41_1.6.5
 12) cray/job/1.5.5-0.1_2.0202.18632.46.1
 13) cray/csa/3.0.0-1_2.0202.18623.63.1
 14) cray/account/1.0.0-2.0202.18612.42.3
 15) cray/projdb/1.0.0-1.0202.18638.45.1
 16) Base-opts/2.2.41A
 17) pgi/9.0.4
 18) xt-libsci/10.4.0
 19) xt-mpt/3.5.0
 20) xt-pe/2.2.41A
 21) xt-asyncpe/3.2
 22) PrgEnv-pgi/2.2.41A
```

Software Environment: `module show pgi`

```
username@jaguarpf-login1:/> module show pgi
-------------------------------------------------------------------
/opt/modulefiles/pgi/9.0.4:

setenv           PGI_VERSION 9.0
setenv           PGI_VERS_STR 9.0.4
setenv           PGI_PATH /opt/pgi/9.0.4
setenv           PGI /opt/pgi/9.0.4
prepend-path     PGROUPD_LICENSE_FILE /opt/pgi/license.dat
prepend-path     PATH /opt/pgi/9.0.4/linux86-64/9.0/bin
prepend-path     MANPATH /opt/pgi/9.0.4/linux86-64/9.0/man
prepend-path     LD_LIBRARY_PATH /opt/pgi/9.0.4/linux86-64/9.0/lib
prepend-path     LD_LIBRARY_PATH /opt/pgi/9.0.4/linux86-64/9.0/libso
module-whatis    PGI compiler for use on XTs
-------------------------------------------------------------------
```

Outline: Compiling

- System Compilers
- Parallel Compiling on Jaguar
- Wrappers and Compiling Tips
- Available Compilers (Serial compilers)
- Useful Compiler Flags

Compiling: System Compilers

The following compilers should be used to build codes on Jaguar!
Use these compilers!

Language	Compiler
C	`cc`
C++	`CC`
Fortran 77, 90 and 95	`ftn`

Note that cc, CC and ftn are actually the Cray XT Series wrappers for invoking the PGI, GNU or Pathscale compilers (discussed later…)

Compiling: Parallel Compiling on Jaguar

- Jaguar has two kinds of nodes:
  - Compute Nodes running the CNL OS
  - Service and login nodes running Linux

- To build a code for the compute nodes, you should use the Cray wrappers `cc`, `CC`, and `ftn`. The wrappers will call the appropriate compiler which will use the appropriate header files and link against the appropriate libraries. Use of wrappers is crucial for building the parallel codes on Cray.

- We highly recommend that the `cc`, `CC`, and `ftn` wrappers be used when building for the compute nodes! Both parallel and serial codes!

- To build a code for the Linux service nodes, you should call the compilers directly.

- We strongly suggest that you don't call the compilers directly if you are building code to run on the compute nodes!

- No long serial jobs should be run on service nodes, use compute nodes instead!

Compiling: Wrappers and Compiling Tips

- Why to use wrappers to build (compile and link) the code:
  - Automatically point to correct compiler based on modules loaded
  - Wrappers automatically find and include paths and libraries of loaded modules (e.g., mpi, libsci)

- Use same makefile for all compilers*

- Calling base compilers directly (e.g., `pgf90`) results in serial code that runs only on login nodes
  - Not what you want! Use wrapper instead and run on compute nodes

* Except compiler-specific flags

Compiling: Available Compilers (Serial compilers)

- Available compilers:
  - Portland Croup (PGI). Module name: PrgEnv-pgi
    - `pgcc`
    - `pgCC`
    - `pgf90/pgf95`
    - `pgf77`
  - GNU. Module name: PrgEnv-gnu
    - `gcc`
    - `g++`
    - `gfortran`
  - Pathscale. Module name: PrgEnv-pathscale
    - `pathcc`
    - `pathCC`
    - `path90/pathf95` (only available if gcc/4.2.1 or higher is loaded)
  - Intel. Module name: PrgEnv-intel
    - `icc` (c/c++ codes)
    - `ifort`
  - Cray compilers. Module name: PrgEnv-cray

Compiling: Default Compilers

- Default compiler is PGI. The list of all packages is obtained by
  - `module avail PrgEnv`

- To use the Cray wrappers with other compilers the programming environment modules need to be swapped, i.e.
  - `module swap PrgEnv-pgi PrgEnv-gnu`
  - `module swap PrgEnv-pgi PrgEnv-cray`

- To just use the GNU/Cray compilers directly load the GNU/Cray module you want:
  - `module load PrgEnv-gnu/2.1.50HD`
  - `module load PrgEnv-cray/1.0.1`

- It is possible to use the GNU compiler versions directly without loading the Cray Programming Environments, but note that the Cray wrappers will probably not work as expected if you do that.

Compiling: Useful Compiler Flags (PGI)

General:

Flag	Comments
`-mp=nonuma`	Compile multithreaded code using OpenMP directives

Debugging:

Flag	Comments
`-g`	For debugging symbols; put first
`-Ktrap=fp`	Trap floating point exceptions
`-Mchkptr`	Checks for unintended dereferencing of null pointers

Optimization:

Flag	Comments
`-fast`	Equivalent to **`-Mvect=sse -Mscalarsse -Mcache_align -Mflushz`**
`-fastsse`	Same as **`-fast`**
`-Mcache_align`	Makes certain that arrays are on cache line boundaries
`-Munroll=c:`*`n`*	Unrolls loops ***n*** times (e.g., ***n***=4)
`-Mipa=fast,inline`	Enables interprocedural analysis (IPA) and inlining, benefits for C++ and Fortran
`-Mconcur`	Automatic parallelization

Compiling: Useful Compiler Flags (GNU)

General:

Flag	Comment
`-fopenmp`	Compile multithreaded code using OpenMP directives

Debugging:

Flag	Comment
`-g`	For debugging symbols; put first
`-finstrument-functions`	For using CrayPat
`-fbounds-check`	Enable generation of runtime checks for array subscripts

Optimization:

Flag	Comments
`-O2 -ffast -math -fomit -frame -pointer -mfpmath=sse`	Recommended first compile/run
`-mfpmath=sse`	Use scalar floating point instructions present in SSE instruction set
`-finline -functions`	Inline simple functions (turned on automatically by **`-O3`**)
`-funroll -loops --param max -unroll -times=`*n*	Unrolls loops ***n*** times (e.g., ***n***=4)

`pathopt2` utility can help identify compiler options that give best optimization

Compiling: Useful Compiler Flags (Pathscale)

General:

Flag	Comments
`-mp`	Compile multithreaded code using OpenMP directives (NOTE: limited support for C++ at this time)

Debugging:

Flag	Comments
`-g`	For debugging symbols; put first
`-LNO:simd_verbose=on`	Get diagnostics
`-trapuv`	Initialize variables to NaN – useful for finding uninitialized variables
`-zerouv`	Initialize variables to 0

Optimization:

Flag	Comments
`-O3 -OPT:Ofast`	Recommended first compile/run
`-OPT:Ofast`	Maximizes performance; generally safe but may impact floating point correctness. Equivalent to `-OPT:ro=2:Olimit=0:div_split=ON:alias=typed`
`-Ofast`	Equivalent to `-O3 -ipa -OPT:Ofast -fno-math-errno`
`-ipa`	Enables interprocedural analysis (IPA) and inlining
`-apo`	Enables autoparallelization

Outline: Running Jobs

Running Jobs: Introduction

- When you log into Jaguar, you are placed on one of the login nodes.
- Login nodes should be used for basic tasks such as file editing, code compilation, data backup, and job submission.
- The login nodes **should not be** used to run production jobs. Production work should be performed on the system's compute resources.
- On Jaguar, access to compute resources is managed by the PBS/TORQUE. Job scheduling and queue management is handled by Moab which interacts with PBS/TORQUE and the XT system software.
- Users either submit the job scripts for batch jobs, or submit a request for interactive job.
- The following pages provide information for getting started with the batch facilities of PBS/TORQUE with Moab as well as basic job execution.

Running Jobs: Glossary

- PBS/TORQUE is an open source resource manager providing control over batch jobs and distributed compute nodes. It is a community effort based on the original PBS project.
- Portable Batch System (or simply PBS) is the computer software that performs job scheduling. Its primary task is to allocate computational tasks, i.e., batch jobs, among the available computing resources. PBS is supported as a job scheduler mechanism by Moab.
- Batch jobs are set up so they can be run to completion without human interaction, so all input data is preselected through scripts or command-line parameters. This is in contrast to "online" or interactive programs which prompt the user for such input.

Running Jobs: Batch Scripts

- Batch scripts can be used to run a set of commands on a systems compute partition.
- The batch script is a shell script containing PBS flags and commands to be interpreted by a shell.
- Batch scripts are submitted to the batch manager, PBS, where they are parsed. Based on the parsed data, PBS places the script in the queue as a job.
- Once the job makes its way through the queue, the script will be executed on the head node of the allocated resources.

Running Jobs: Example Batch Script

```
1: #!/bin/bash
2: #PBS -A XXXYYY
3: #PBS -N test
4: #PBS -j oe
5: #PBS -l walltime=1:00:00,size=192
6:
7: cd $PBS_O_WORKDIR
8: date
9: aprun -n 192 ./a.out
```

NOTE: Since users cannot share nodes, size requests must be
- Шa multiple of 4 on the XT4 or
- Шa multiple of 12 on the XT5.

This batch script can be broken down into the following sections:

- Shell interpreter
  - Line 1
  - Can be used to specify an interpreting shell.
- PBS commands
  - The PBS options will be read and used by PBS upon submission.
  - Lines 2–5
    - 2: The job will be charged to the XXXYYY project.
    - 3: The job will be named “test.”
    - 4: The jobs standard output and error will be combined.
    - 5: The job will request 192 cores for 1 hour.
  - Please see the PBS Options page for more options.
- Shell commands
  - Once the requested resources have been allocated, the shell commands will be executed on the allocated nodes head node.
  - Lines 6–9
    - 6: This line is left blank, so it will be ignored.
    - 7: This command will change directory into the script's submission directory. We assume here that the job was submitted from a directory in /lustre/scratch/.
    - 8: This command will run the date command.
    - 9: This command will run the executable a.out on 192 cores with a.out.

Running Jobs: Submitting Batch Jobs - `qsub`

- All job resource management handled by Torque.
- Batch scripts can be submitted for execution using the `qsub` command.
- For example, the following will submit the batch script named `test.pbs`:

```
qsub test.pbs
```

- If successfully submitted, a PBS job ID will be returned. This ID can be used to track the job.

Running Jobs: Interactive Batch Jobs

- Batch scripts are useful for submitting a group of commands, allowing them to run through the queue, then viewing the results. It is also often useful to run a job interactively. However, users are not allowed to directly run on compute resources from the login module. Instead, users must use a batch-interactive PBS job. This is done by using the `-I` option to `qsub`.
- For interactive batch jobs, PBS options are passed through `qsub` on the command line:

```
qsub -I -A XXXYYY -q debug -V -l size=24,walltime=1:00:00
```

This request will…

`-I`	Start an interactive session
`-A`	Charge to the "XXXYYY" project
`-q debug`	Run in the debug queue
`-V`	Import the submitting users environment
`-l size=24,walltime=1:00:00`	Request 24 compute cores for one hour

Running Jobs: PBS Options

Necessary PBS options:

Option	Use	Description
A	`#PBS -A <account>`	Causes the job time to be charged to <account>. The account string XXXYYY is typically composed of three letters followed by three digits and optionally followed by a subproject identifier. The utility showusage can be used to list your valid assigned project ID(s). This is the only option required by all jobs.
l	`#PBS -l size=<cores>`	Maximum number of compute cores. Must request an entire node (multiples of 4 on the XT4, and 12 on the XT5).
	`#PBS -l walltime=<time>`	Maximum wall-clock time. <time> is in the format HH:MM:SS. Default is 45 minutes.

Running Jobs: PBS Options (cont.)

Commonly used, but not necessary PBS Options:

Option	Use	Description
`l`	`#PBS -l feature=<target>`	Run only on the specified target. Currently the available target is XT5 with 1 or 2 GB of memory per node. The default is to run on the first available. It is recommended to use the default. The other option is to specify "2gbpercore" to run on 16 GB nodes only.
`o`	`#PBS -o <name>`	Writes standard output to <name> instead of <job script>.o$PBS_JOBID. $PBS_JOBID is an environment variable created by PBS that contains the PBS job identifier.
`e`	`#PBS -e <name>`	Writes standard error to <name> instead of <job script>.e$PBS_JOBID.
`j`	`#PBS -j {oe,eo}`	Combines standard output and standard error into the standard error file (eo) or the standard out file (oe).
`m`	`#PBS -m a`	Sends email to the submitter when the job aborts.
	`#PBS -m b`	Sends email to the submitter when the job begins.
	`#PBS -m e`	Sends email to the submitter when the job ends.
`M`	`#PBS -M <address>`	Specifies email address to use for -m options.
`N`	`#PBS -N <name>`	Sets the job name to <name> instead of the name of the job script.
`S`	`#PBS -S <shell>`	Sets the shell to interpret the job script.
`q`	`#PBS -q <queue>`	Directs the job to the specified queue.This option is not required to run in the general production queue.
`V`	`#PBS -V`	Exports all environment variables from the submitting shell into the batch shell.

Running Jobs: PBS Environment Variables

- `PBS_O_WORKDIR`
  - PBS sets the environment variable PBS_O_WORKDIR to the directory where the batch job was submitted.
  - By default, a job starts in your home directory.
  - Include the following command in your script if you want it to start in the submission directory:

```
cd $PBS_O_WORKDIR
```

- `PBS_JOBID`
  - PBS sets the environment variable PBS_JOBID to the job's ID.
  - A common use for PBS_JOBID is to append the job's ID to the standard output and error file(s), such as the following:

```
PBS -o scriptname.o$PBS_JOBID
```

- `PBS_NNODES`
  - PBS sets the environment variable PBS_NNODES to the number of cores requested. This means that number of nodes requested on a 12-core architecture would be $PBS_NNODES/12.

Running Jobs: Altering Batch Jobs – `qdel,qhold,qrls`

- Command: `qdel`
  - Jobs in the queue in any state can be stopped and removed from the queue using the command qdel.
  - For example, to remove a job with a PBS ID of 1234, use the following command: `qdel 1234`

- Command: `qhold`
  - Jobs in the queue in a non-running state may be placed on hold using the qhold command. Jobs placed on hold will not be removed from the queue, but they will not be eligible for execution.
  - For example, to move a currently queued job with a PBS ID of 1234 to a hold state, use the following command: `qhold 1234`

- Command: `qrls`
  - Once on hold the job will not be eligible to run until it is released to return to a queued state. The qrls command can be used to remove a job from the held state.
  - For example, to release job 1234 from a held state, use the following command: `qrls 1234`

Running Jobs: Altering Batch Jobs – `qalter`

- Command: `qalter`
  - Non-running jobs in the queue can be modified with the PBS `qalter` command. **Please note: Jobs in a running state cannot be altered.**
  - The following uses job 130494 as an example:

```
> qstat -a 130494nid03588: ORNL/CCS
                                                    Req'd  Req'd  Elap
Job ID Username Queue Jobname SessID NDS Tasks Memory Time  S Time
------ -------- ----- ------- ------ --- ----- ------ ---- - ----

130494.nid03588 cfuson debug t1 -- 1 4300 0b 00:10 Q --
>
```

- Modify the job's name
  - `qalter -N newname 130494`
- Modify the number of requested cores
  - `qalter -l size=4800 130494`
- Modify the job's wall time
  - `qalter -l walltime=01:00:00 130494`

Running Jobs: Monitoring Job Status - `qstat`

PBS and Moab provide multiple tools to view queue, system, and job statuses.
Command: `qstat`
Use `qstat -a` to check the status of submitted jobs:

```
nid00004:
                                                          Req'd  Req'd     Elap
Job ID Username Queue Jobname SessID NDS Tasks Memory Time   S Time
------ -------- ----- ------- ------ --- ----- ------ ----- - -----
29668   user1   batch job2    21909    1  256   --     08:00 R 02:28
29894   user2   batch run128   --      1  128   --     02:30 Q --
29895   user3   batch STDIN   15921    1  1     --     01:00 R 00:10
29896   user2   batch jobL    21988    1  2048  --     01:00 R 00:09
29897   user4   debug STDIN   22367    1  2     --     00:30 R 00:06
29898   user1   batch job1    25188    1  1     --     01:10 C 00:00
```

Job ID	PBS assigned job ID.
Username	Submitting user's user ID.
Queue	Queue into which the job has been submitted.
Jobname	PBS job name. This is given by the PBS -n option in the PBS batch script. Or, if the -n option is not used, PBS will use the name of the batch script.
SessID	Associated session ID.
NDS	PBS node count. Not accurate; will be one.
Tasks	Number of cores requested by the job's -size option.
Req'd Memory	Job's requested memory.
Req'd Time	Job's given wall time.
S	Job's current status. See the status listings below.
Elap Time	Job's time spent in a running status. If a job is not currently or has not been in a run state, the field will be blank.

Status Value	**Meaning**
E	Exiting after having run
H	Held
Q	Queued; eligible to run
R	Running
S	Suspended
T	Being moved to new location
W	Waiting for its execution time
C	Recently completed (within the last 5 minutes)

Running Jobs: `showq, checkjob`

Command : `showq`

The Moab utility `showq` gives a more detailed description of the queue and displays it in the following states:

Active These jobs are currently running.

Eligible These jobs are currently queued awaiting resources. A user is allowed five jobs in the eligible state.

Blocked These jobs are currently queued but are not eligible to run. Common reasons for jobs in this state are jobs on hold, the owning user currently having five jobs in the eligible state, and running jobs in the longsmall queue.

Command : `checkjob`

The Moab utility `checkjob` can be used to view details of a job in the queue.

For example, if job 736 is a job currently in the queue in a blocked state, the following can be used to view why the job is in a blocked state:

`checkjob 736` The return may contain a line similar to the following:

```
BlockMsg: job 736 violates idle HARD MAXJOB limit of 2 for
      user (Req: 1 In Use: 2)
```

This line indicates the job is in the blocked state because the owning user has reached the limit of two jobs currently in the eligible state.

Running Jobs: showstart, showbf, xtprocadmin

Command : `showstart`

The Moab utility `showstart` gives an estimate of when the job will start.

```
showstart 100315
job 100315 requires 16384 procs for 00:40:00
Estimated Rsv based start in 15:26:41 on Fri Sep 26 23:41:12
Estimated Rsv based completion in 16:06:41 on Sat Sep 27 00:21:12
```

Since the start time may change dramatically as new jobs with higher priority are submitted, so you need to periodically rerun the command.

Command : `showbf`

The Moab utility `showbf` gives the current backfill. This can help to build a job which can be backfilled immediately. As such, it is primarily useful for small jobs.

Command : `xtprocadmin`

The utility `xtprocadmin` can be used to see what jobs are currently running and which nodes they are running on.

Running Jobs: Job Execution - `aprun`

- By default, commands will be executed on the job's associated service node.
- The `aprun` command is used to execute a job on one or more compute nodes.
- The XT's layout should be kept in mind when running a job using `aprun`. The XT5 partition currently contains two hex-core processors (a total of 12 cores) per compute node. While the XT4 partition currently contains one quad-core processor (a total of 4 cores) per compute node.
- The `PBS size` option requests compute cores.

Running Jobs: Job Execution – Service Node

- The PBS script is executed on the aprun node (or login node for interactive jobs).
- If executables are called directly (*eg* ./a.out), they will be run serially on the service node. This may be useful for records keeping, staging data, *etc*.
- Please run any memory- or computationally-intensive programs using `aprun`, otherwise it bogs down the node, and may cause system problems.
- You may run non-MPI (serial) programs on a compute node using `aprun` (discussed later).

Running Jobs: Basic `aprun` options

Option	Description
`-D`	Debug (shows the layout aprun will use)
`-n`	Number of MPI tasks. Note: If you do not specify the number of tasks to aprun, the system will default to 1.
`-N`	Number of tasks per Node. (XT5: 1 – 12) and (XT4: 1 – 4) NOTE: Recall that the XT5 has two Opterons per compute node. On the XT5, to place one task per quad-core Opteron, use -S 1 (not -N 1 as on the XT4). On the XT4, because there is only one Opteron per node, the -S 1 and -N1 will result in the same layout.
`-m`	Memory required per task. Default: 4-core, 8-GB Cray XT4 nodes (8 GB / 4 CPUs = 2 GB) XT4: A maximum of 2GB per core; 2.1GB will allocate two cores for the task
`-d`	Number of threads per MPI task. Note: As of CLE 2.1, this option is very important. If you specify OMP_NUM_THREADS but do not give a -d option, aprun will allocate your threads to a single core. You must use OMP_NUM_THREADS to specify the number of threads per MPI task, and you must use -d to tell aprun how to place those threads.
`-S`	Number of PEs to allocate per NUMA node.
`-ss`	Strict memory containment per NUMA node.

Running Jobs: XT5 example

`aprun –n 24 ./a.out` will run a.out across 24 cores. This requires two compute nodes. The MPI task layout would be as follows:

Compute Node 1

Opteron 0						Opteron 1					
Core 0	**Core 1**	**Core 2**	**Core 3**	**Core 4**	**Core 5**	**Core 0**	**Core 1**	**Core 2**	**Core 3**	**Core 4**	**Core 5**
0	1	2	3	4	5	6	7	8	9	10	11

Compute Node 2

Opteron 0						Opteron 1					
Core 0	**Core 1**	**Core 2**	**Core 3**	**Core 4**	**Core 5**	**Core 0**	**Core 1**	**Core 2**	**Core 3**	**Core 4**	**Core 5**
12	13	14	15	16	17	18	19	20	21	22	23

The following will place tasks in a round robin fashion.

```
> setenv MPICH_RANK_REORDER_METHOD 0
> aprun -n 24 a.out
```

```
Rank 0,  Node 1, Opteron 0, Core 0
Rank 1,  Node 2, Opteron 0, Core 0
Rank 2,  Node 1, Opteron 0, Core 1
Rank 3,  Node 2, Opteron 0, Core 1
Rank 4,  Node 1, Opteron 0, Core 2
Rank 5,  Node 2, Opteron 0, Core 2
Rank 6,  Node 1, Opteron 0, Core 3
Rank 7,  Node 2, Opteron 0, Core 3
Rank 8,  Node 1, Opteron 0, Core 4
Rank 9,  Node 2, Opteron 0, Core 4
Rank 10, Node 1, Opteron 0, Core 5
Rank 11, Node 2, Opteron 0, Core 5
```

```
Rank 0,  Node 1, Opteron 1, Core 0
Rank 1,  Node 2, Opteron 1, Core 0
Rank 2,  Node 1, Opteron 1, Core 1
Rank 3,  Node 2, Opteron 1, Core 1
Rank 4,  Node 1, Opteron 1, Core 2
Rank 5,  Node 2, Opteron 1, Core 2
Rank 6,  Node 1, Opteron 1, Core 3
Rank 7,  Node 2, Opteron 1, Core 3
Rank 8,  Node 1, Opteron 1, Core 4
Rank 9,  Node 2, Opteron 1, Core 4
Rank 10, Node 1, Opteron 1, Core 5
Rank 11, Node 2, Opteron 1, Core 5
```

Running Jobs: XT4 example

`aprun -n8 a.out` will run the MPI executable a.out on a total of eight cores, four cores on two compute nodes. The MPI tasks will be allocated in the following sequential fashion:

Compute Node 1

Opteron 0

Core 0	**Core 1**	**Core 2**	**Core 3**
0	1	2	3

Compute Node 2

Opteron 0

Core 0	**Core 1**	**Core 2**	**Core 3**
0	1	2	3

The following will place tasks in a round robin fashion.

```
> setenv MPICH_RANK_REORDER_METHOD 0
> aprun -n 8 a.out
Rank 0,  Node 1, Opteron 0, Core 0
Rank 1,  Node 2, Opteron 0, Core 0
Rank 2,  Node 1, Opteron 0, Core 1
Rank 3,  Node 2, Opteron 0, Core 1
Rank 4,  Node 1, Opteron 0, Core 2
Rank 5,  Node 2, Opteron 0, Core 2
Rank 6,  Node 1, Opteron 0, Core 3
Rank 7,  Node 2, Opteron 0, Core 3
```

Running Jobs: Memory Affinity

- Each Opteron CPU on a node and its memory is organized into a NUMA node.
- Memory Affinity - each XT5 (XT4) node contains two (one) NUMA nodes.
- Applications may use resources from one or both NUMA nodes. The following `aprun` options allow control of application NUMA node use.

`-S pes_per_numa_node`

Number of PEs to allocate per NUMA node, *`pes_per_numa_node`* can be 1, 2, 3, or 4.

`-ss`

is the option for strict memory containment per NUMA node.

- `-ss` option: The default is to allow remote NUMA node memory access. This option prevents memory access of the remote NUMA node. Because the XT4 has only one NUMA node per node, this option does not apply to the XT4.

XT5 Example:
The following will run a.out on 4 cores, one core per NUMA node.

```
aprun -n 4 -S 1 a.out

Rank 0, Node 0, Opteron 0, Core 0
Rank 1, Node 0, Opteron 1, Core 0
Rank 2, Node 1, Opteron 0, Core 0
Rank 3, Node 1, Opteron 1, Core 0
```

XT4 Example:
The following will run a.out on 4 cores, all will be on one NUMA node.

```
aprun -n 4 -S 4 a.out

Rank 0, Node 0, Opteron 0, Core 0
Rank 1, Node 0, Opteron 0, Core 1
Rank 2, Node 0, Opteron 0, Core 2
Rank 3, Node 0, Opteron 0, Core 3
```

Running Jobs: Threads

- The system supports threaded programming within a compute node.
- On the XT5, threads may span both Opterons within a single compute node, but cannot span compute nodes.
- Users have a great deal of flexibility in thread placement. Several examples are shown below.
- Note: Under CNL 2.1, threaded codes must use the

  `aprun -d depth` option

  The -d option specifies the number of threads per task. Without the option all threads will be started on the same core. Under previous CNL versions the option was not required. The number of cores used is calculated by multiplying the value of -d by the value of -n.
- Focus of this discussion will be OpenMP threads

Running Jobs: Threads – XT5 Example 1

These examples are written for bash. If using csh/tcsh, you should change the `export OMP_NUM_THREADS = x` lines to `setenv OMP_NUM_THREADS x`

Example 1: Launch 2 MPI tasks, each with 12 threads (this requests 2 compute nodes and requires a size request of 24):

```
export OMP_NUM_THREADS=12
> aprun -n2 -d12 a.out
Rank 0, Thread 0, Node 0, Opteron 0, Core 0 <-- MASTER
Rank 0, Thread 1, Node 0, Opteron 0, Core 1 <-- slave
Rank 0, Thread 2, Node 0, Opteron 0, Core 2 <-- slave
Rank 0, Thread 3, Node 0, Opteron 0, Core 3 <-- slave
Rank 0, Thread 4, Node 0, Opteron 0, Core 4 <-- slave
Rank 0, Thread 5, Node 0, Opteron 0, Core 5 <-- slave
Rank 0, Thread 6, Node 0, Opteron 1, Core 6 <-- slave
Rank 0, Thread 7, Node 0, Opteron 1, Core 1 <-- slave
Rank 0, Thread 8, Node 0, Opteron 1, Core 2 <-- slave
Rank 0, Thread 9, Node 0, Opteron 1, Core 3 <-- slave
Rank 0, Thread 10,Node 0, Opteron 1, Core 4 <-- slave
Rank 0, Thread 11,Node 0, Opteron 1, Core 5 <-- slave
Rank 1, Thread 0, Node 1, Opteron 0, Core 0 <-- MASTER
Rank 1, Thread 1, Node 1, Opteron 0, Core 1 <-- slave
Rank 1, Thread 2, Node 1, Opteron 0, Core 2 <-- slave
-----//--------//------//---------//------//---------
Rank 1, Thread 9, Node 1, Opteron 1, Core 3 <-- slave
Rank 1, Thread 10,Node 1, Opteron 1, Core 4 <-- slave
Rank 1, Thread 11,Node 1, Opteron 1, Core 5 <-- slave
```

Running Jobs: Threads – XT5 Example 2

- Example 2: Launch 4 MPI tasks, each with 6 threads. Place 1 MPI task per Opteron (this requests 2 compute nodes and requires a size request of 24):

```
export OMP_NUM_THREADS=6
> aprun -n4 -d6 -S1 a.out
Rank 0, Thread 0, Node 0, Opteron 0, Core 0 <-- MASTER
Rank 0, Thread 1, Node 0, Opteron 0, Core 1 <-- slave
Rank 0, Thread 2, Node 0, Opteron 0, Core 2 <-- slave
Rank 0, Thread 3, Node 0, Opteron 0, Core 3 <-- slave
Rank 0, Thread 4, Node 0, Opteron 0, Core 4 <-- slave
Rank 0, Thread 5, Node 0, Opteron 0, Core 5 <-- slave
Rank 1, Thread 0, Node 0, Opteron 1, Core 0 <-- MASTER
Rank 1, Thread 1, Node 0, Opteron 1, Core 1 <-- slave
Rank 1, Thread 2, Node 0, Opteron 1, Core 2 <-- slave
Rank 1, Thread 3, Node 0, Opteron 1, Core 3 <-- slave
Rank 1, Thread 4, Node 0, Opteron 1, Core 4 <-- slave
Rank 1, Thread 5, Node 0, Opteron 1, Core 5 <-- slave
Rank 2, Thread 0, Node 1, Opteron 0, Core 0 <-- MASTER
Rank 2, Thread 1, Node 1, Opteron 0, Core 1 <-- slave
Rank 2, Thread 2, Node 1, Opteron 0, Core 2 <-- slave
Rank 2, Thread 3, Node 1, Opteron 0, Core 3 <-- slave
Rank 2, Thread 4, Node 1, Opteron 0, Core 4 <-- slave
Rank 2, Thread 5, Node 1, Opteron 0, Core 5 <-- slave
Rank 3, Thread 0, Node 1, Opteron 1, Core 0 <-- MASTER
Rank 3, Thread 1, Node 1, Opteron 1, Core 1 <-- slave
Rank 3, Thread 2, Node 1, Opteron 1, Core 2 <-- slave
Rank 3, Thread 3, Node 1, Opteron 1, Core 3 <-- slave
Rank 3, Thread 4, Node 1, Opteron 1, Core 4 <-- slave
Rank 3, Thread 5, Node 1, Opteron 1, Core 5 <-- slave
```

Running Jobs: Threads – XT5 Example 3

- Example 3: Launch 4 MPI tasks, each with 2 threads. Only place 1 MPI task (its two threads) on each Opteron. (This requests 2 compute nodes and requires a size request of 24 even though only 8 cores are actually being used):

```
export OMP_NUM_THREADS=2
> aprun -n4 -d2 -S1 a.out
Rank 0, Thread 0, Node 0, Opteron 0, Core 0 <-- MASTER
Rank 0, Thread 1, Node 0, Opteron 0, Core 1 <-- slave
Rank 1, Thread 0, Node 0, Opteron 1, Core 0 <-- MASTER
Rank 1, Thread 1, Node 0, Opteron 1, Core 1 <-- slave
Rank 2, Thread 0, Node 1, Opteron 0, Core 0 <-- MASTER
Rank 2, Thread 1, Node 1, Opteron 0, Core 1 <-- slave
Rank 3, Thread 0, Node 1, Opteron 1, Core 0 <-- MASTER
Rank 3, Thread 1, Node 1, Opteron 1, Core 1 <-- slave
```

Running Jobs: Threads – XT4 Example 1

- Example 1: Launch 2 MPI tasks, each with 4 threads (this requests 2 compute nodes and requires a size request of 8):

```
export OMP_NUM_THREADS=4
> aprun -n2 -d4 a.out
Rank 0, Thread 0, Node 0, Opteron 0, Core 0 <-- MASTER
Rank 0, Thread 1, Node 0, Opteron 0, Core 1 <-- slave
Rank 0, Thread 2, Node 0, Opteron 0, Core 2 <-- slave
Rank 0, Thread 3, Node 0, Opteron 0, Core 3 <-- slave
Rank 1, Thread 0, Node 1, Opteron 0, Core 0 <-- MASTER
Rank 1, Thread 1, Node 1, Opteron 0, Core 1 <-- slave
Rank 1, Thread 2, Node 1, Opteron 0, Core 2 <-- slave
Rank 1, Thread 3, Node 1, Opteron 0, Core 3 <-- slave
```

Running Jobs: Threads – XT4 Example 2

- Example 2: Launch 2 MPI tasks, each with 2 threads. Place 1 MPI task per Opteron (this requests 2 compute nodes and requires a size request of 8):

```
export OMP_NUM_THREADS=4
> aprun -n2 -d2 -S1 a.out
Rank 0, Thread 0, Node 0, Opteron 0, Core 0 <-- MASTER
Rank 0, Thread 1, Node 0, Opteron 0, Core 1 <-- slave
Rank 2, Thread 0, Node 1, Opteron 0, Core 0 <-- MASTER
Rank 2, Thread 1, Node 1, Opteron 0, Core 1 <-- slave
```

Running Jobs: MPI Task Layout

The default MPI task layout is *SMP-style*. This means MPI will sequentially allocate all cores on one node before allocating tasks to another node.

Changing/Viewing Layout

- The layout order can be changed using the environment variable `MPICH_RANK_REORDER_METHOD`. See man `intro_mpi` for more information.

- Task layout can be seen by setting `MPICH_RANK_REORDER_DISPLAY` to 1.

Running Jobs: Single-Processor (Serial) Jobs

- Serial programs which are memory or computationally intensive should never be run on the service nodes (anything outside of `aprun`).
- Service nodes have limited resources shared between all users. When they run out the system problems may result.
- To run serial programs on the compute nodes, the program must be compiled with the compiler wrappers: `cc`, `CC` or `ftn`. You would then need to request one socket (12 cores) with PBS (`#PBS -l size=12`).
- Use the following line to run a serial executable on a compute node:

```
aprun -n 1 ./a.out
```

- Note that on both XT4 and XT5 running a serial job will give you an access to all memory of the socket – 8 Gb.
- Running a serial job on a single core will occupy the whole node, so that the remaining cores (three cores for XT4 node and seven cores for XT5 node) will be idling.
- The following slide shows you how to make use of these idling nodes by running several copies of serial job on them.

Running Jobs: Running Multiple Single-Processor Programs

The following batch script shows how to run multiple copies of a serial program on a compute node:

```
#!/bin/csh
#PBS -A TG-XXXXXXXXX
#PBS -N run_serial
#PBS -l walltime=00:30:00,size=12
#PBS -j oe
#PBS -V

set echo
cd /lustre/scratch/$USER/serial_job

# Use aprun to start a shell script which runs 12 copies of the
# of the same executable on a compute node
# Note: all aprun options specified below are required
# -n 1 # run on a single node
# -d 12 # allows the script to access all the memory on the node
# -cc none  # allows each serial process to run on its own core
# -a xt # required by aprun to run a script instead of a program

aprun -n 1 -d 12 -cc none -a xt ./run_serial
```

Running Jobs: Running Multiple Single-Processor Programs

The run_serial script looks like this:

```
#!/bin/sh # This must be /bin/sh (other shells do not
work)

# Run 12 copies of serial_code in the background
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &
./serial_code &

# Wait until all copies of serial_code have finished wait
```

Third-Party Software

- NCCS has installed many third-party software packages, libraries, etc., and created module files for them
  - Third-party applications (e.g., MATLAB, GAMESS)
  - Latest versions or old versions not supported by vendor (e.g., fftw/3.1.2)
  - Suboptimal versions to do proof-of-concept work (e.g., blas/ref)
  - Debug versions (e.g., petsc/2.3.3-debug)
- NCCS modules available via `module load` command, installed in `/sw/xt/` directory

Outline: Resources for Users

Resources for Users: Getting Started

- About Jaguar

  http://www.nccs.gov/computing-resources/jaguar/

- Quad Core AMD Opteron Processor Overview

  http://www.nccs.gov/wp-content/uploads/2008/04/amd_craywkshp_apr2008.pdf

- PGI Compilers for XT5

  http://www.nccs.gov/wp-content/uploads/2008/04/compilers.ppt

- NCCS Training & Education – archives of NCCS workshops and seminar series, HPC/ parallel computing references

  http://www.nccs.gov/user-support/training-education/

- 2009 Cray XT5 Quad-core Workshop

  http://www.nccs.gov/user-support/training-education/workshops/2008-cray-xt5-quad-core-workshop/

Resources for Users: Advanced Topics

- Debugging Applications Using TotalView

  http://www.nccs.gov/user-support/general-support/software/totalview

- Debugging Applications Using DDT

  http://www.nccs.gov/computing-resources/jaguar/software/?software=ddt

- Using Cray Performance Tools - CrayPat

  http://www.nccs.gov/computing-resources/jaguar/debugging-optimization/cray-pat/

- I/O Tips for Cray XT4

  http://www.nccs.gov/computing-resources/jaguar/debugging-optimization/io-tips/

- NCCS Software

  http://www.nccs.gov/computing-resources/jaguar/software/

Resources for Users: More Information

- NCCS website

  http://www.nccs.gov/

- Cray Documentation

  http://docs.cray.com/

- Contact us

  help@nccs.gov

Go to Menu